

DOCUMENT RESUME

ED 209 905

EC 140 625

AUTHOR Garrison, Wayne M.; Coggiola, Deborah C.
TITLE Time Limits in Standardized Testing: Effects on Ability Estimation.
INSTITUTION National Technical Inst. for the Deaf, Rochester, N. Y.
SPONS AGENCY Department of Health, Education, and Welfare, Washington, D.C.
REPORT NO NTID-PS-37
PUB DATE Sep 80
NOTE 17p.; For related documents, see EC 140 611-630.
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Aptitude Tests; College Students; *Deafness; Higher Education; *Standardized Tests; Student Evaluation; *Timed Tests
IDENTIFIERS *Differential Aptitude Test

ABSTRACT

The study involving 614 hearing impaired students investigated the adequacy of time limits currently followed in the administration of two subtests of the Differential Aptitude Test battery when used with young deaf students. Thirty-six students who did not complete the Space Relations portion of the battery under the timed condition, and 11 who did not complete the Abstract Reasoning subtest within the time allowed were instructed to complete the test on a separate answer sheet. Examinees' initial and revised measurements were compared for statistical equivalence in terms of the standard error of measurement computed from independent sample data. No evidence was adduced to indicate that deaf examinees would profit from an extension of testing time. (Author)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED209905

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TIME LIMITS IN STANDARDIZED TESTING:
EFFECTS ON ABILITY ESTIMATION

Paper Series No. 37
September, 1980

Wayne M. Garrison
Deborah C. Coggiola

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

NTID

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

NTID's principal goal in doing research is to influence the education, training and career placement of deaf citizens through systematic examination of issues related to deafness. As one part of NTID's total research effort, the Department of Research and Development conducts descriptive and experimental research. Research findings are used in the development of programs and materials in the areas of learning and instruction, personal and social growth, and career development of deaf students. This document was developed in the course of an agreement with the U.S. Department of Health, Education and Welfare.

Abstract

This study investigated the "adequacy" of time limits currently followed in the administration of two subtests of the Differential Aptitude Test battery when used with young deaf adults. Thirty-six students who did not complete the Space Relations portion of the battery under the timed condition, and 11 who did not complete the Abstract Reasoning subtest within the time allowed, were instructed to complete the test on a separate answer sheet. Examinees' initial and revised measurements were compared for statistical equivalence in terms of the standard error of measurement computed from independent sample data. No evidence was adduced to indicate that deaf examinees would profit from an extension of testing time.

Time Limits in Standardized Testing: Effects on Ability Estimation

When standardized mental tests are administered individually, it is not always necessary that the examinee attempt every item included in a test. Many tests of mental ability are constructed such that the items are arranged in order of increasing difficulty. Consequently, the examiner begins testing the subject at a point in the difficulty sequence where it seems likely that the subject will demonstrate some number of consecutive successes and, through successive movements up and down the scale of difficulty, the subject's difficulty ceiling (i.e., ability) is determined. Such individualized procedures make testing efficient, spare subjects from the boredom of too many easy items and the frustration of too many very difficult items, and ensure that most of the testing period will be spent on those problems in the test that are most discriminating at the particular respondent's level of ability (Jensen, 1980).

In group testing, the desirable aspects of individualized testing procedures are not operative. All examinees must start at the beginning of a test. How far a subject will progress through the test will depend upon the number and difficulty of the items, and the time allowed. Many group tests in current use are timed but, presumably, they allow enough time for the majority of examinees to reach their individual ceilings of difficulty. However, the time limits governing the administration of standardized mental tests typically were derived from work with other than handicapped or minority populations.

In general, research evidence has indicated that the condition of timed versus untimed tests of mental ability is a negligible factor in the observed mean test score differences between various racial and socioeconomic groups (Jensen, 1980). That is, when the items in a test are evenly graded in difficulty and the test has sufficient "top" (i.e., very difficult items), and faster students can finish the test, but reach their

difficulty ceiling before the end of the test, then giving examinees additional time beyond the prescribed limits adds little to examinees' scores and has negligible effects upon the rank order of respondents' scores (Cronbach, 1960).

Because the findings of the research summarized above were not derived from test data on examinees with sensory impairments, and as a further test of the extent to which standardized measures administered to deaf respondents meet standards of validity, the present study examined the performance of hearing-impaired students on two subtests of the Differential Aptitude Test battery (DAT: Bennett, Seashore & Wesman, 1966) under strictly timed and "allow to finish" conditions. Specifically, the present study was conducted to evaluate the adequacy of time limits followed in the administration of the Space Relations and Abstract Reasoning DAT subtests when the respondents are young deaf adults.

METHOD

Data Source

Data for the present study were obtained from 306 (Space Relations test) and 308 (Abstract Reasoning test) hearing-impaired students tested during the 1980 Summer Vestibule Program. Students were tested in groups of approximately 12-14 persons each by their respective Career Planning Seminar (CPS) counselors. Testing occurred during the first week of the orientation program.

Instrumentation and Procedure

The DAT subtests were administered to all students under the time limits specified by the test publisher. The Abstract Reasoning subtest consists of 50 items, is nonverbal, and is intended to assess an individual's reasoning ability. The Space Relations subtest also is nonverbal, consists of 60 items, and is intended to measure the ability to deal with

concrete material through visualization. Both subtests utilize a "spiral omnibus" pattern, wherein test items are steeply graded in difficulty.

Students participating in the orientation program were distributed into 24 CPS groups. For six of the CPS groups, examinees who had not completed the Abstract Reasoning test during the time allotted were instructed to complete the test on a separate answer sheet (immediately following the time limit expiration). Six other CPS groups were instructed likewise for the Space Relations test. Separate answer sheets were used in both instances to guard against the possibility of students revising responses to items attempted under the timed condition.

RESULTS AND DISCUSSION

Space Relations Test (SRE)

Of the 306 individuals tested with the SRE, 170 (55.6%) responded to all items (60 total) and completed the test within the 25-minute time limit. For those individuals not completing the test within the time allowed ($n = 136$ total or 44.4% of the study sample), the mean number of items for which responses were not provided was approximately ten (range 1-35).

For purposes of analyses to be discussed below, the study sample was divided into three groups. Group A consisted of the 170 persons who had responded to all items and completed the test. The 136 individuals remaining were divided into two groups. Group B consisted of 100 examinees who had not completed the SRE within the time limit, and who were not given the opportunity to complete the test. Group C consisted of 36 individuals who initially did not complete the test within the time limit, but were allowed to finish.

Table 1 provides descriptive statistics bearing on the test performance of the three groups of individuals. Simple inspection of the test score means and standard deviations for groups B and C suggests no significant difference in test performance and, therefore,

Table 1
Test Performance of Three Groups of Examinees (SRE)

	Group A	Group B	Group C
n	170	100	36
$\bar{x}$	38.69	35.88	35.14
sd	12.22	10.47	10.40
$s_{\bar{x}}$	.94	1.05	1.73

also suggests that the 36 individuals included in the "allow to finish" condition were representative of the larger number of individuals who did not complete the test, and were not permitted to do so. A statistical test of the difference between the means of the two groups yielded a t-statistic of 0.37 (n.s.).¹

A comparison of the test score means of groups A and B yielded a t-statistic of 1.99 ($p < .05$), indicating a statistically significant difference. This finding, specifically the higher mean test performance of group A respondents relative to that of group B respondents, leads to two alternative hypotheses concerning the nature of group differences: (1) group A examinees were more able than group B examinees, respondent ability being a determinant in test completion; and (2) the observed difference in test performance is due to non-ability related factors (e.g., more instances of chance success in the test records of group A respondents compared with those of group B respondents, owing to the greater number (also difficulty) of test items attempted by members of group A).

To examine the tenability of these alternative hypotheses, the test score means of groups A and B were compared again, basing test scores only on the first 50 of the 60 items comprising the test. Recall that respondents in group A completed, on the average, approximately ten items more than did respondents in group B, making the present comparison freer of the effects of "time" than was true for the former comparison. For group A, the mean number of items correctly answered (based upon 50 items) was 35.21, with standard deviation equal to 10.21. For group B, the corresponding mean and standard deviation were found to be 34.59 and 9.59, respectively. A test of the difference between these two means yielded a t-statistic of 0.50 (n.s.). With this new information, the previous finding of a statistically significant difference in the mean performance of

groups A and B (based upon all 60 SRE items) can be attributed to group A's observed successes (and group B's unknown successes) on the last ten items in the SRE.

Now, because the SRE follows a spiral omnibus pattern, with the last ten items in the test being the most difficult, it is useful at this point to examine the increase in the mean test performance of group A between the 50 and 60 item sets in the light of the chance success factor. That is, to the extent that the difference between the means observed for group A for 50 and 60 items, respectively, is overshadowed by the difference which would be expected from chance success, the hypothesis of "true" score gains between the two conditions becomes untenable.

For group A respondents, the difference between the means observed under the 50 and 60 item conditions was 3.48. With a difference of ten items between the conditions, and four response alternatives per item, a difference of $10 \div (4-1)$, or 3.33, would be predicted based upon a hypothesis of chance success (see Stanley & Hopkins, 1972), very close to the 3.48 difference actually observed. In the light of this finding, suggesting that group A respondents had reached their difficulty ceilings within the first 50 items of the test, a correction for chance scoring formula was applied to all respondents' test data.²

The use of correction for chance formulae is justified when the number of omitted test items varies appreciably among examinees (Stanley & Hopkins, 1972), as observed in the present study. Application of formula scoring to a set of test data has the effect of: (1) reducing the scores of respondents who resort to guessing, while not affecting the scores of more able respondents; (2) increasing the variability of possible scores; and (3) in the present situation, making possible test score change in both the positive and negative direction.

Table 2 provides descriptive statistics bearing on the test performance of the three groups of respondents after the correction for chance formula was applied to all individuals' test data. Note the striking similarity between the test score means of groups B and C. Moreover, a test of the difference between the means of groups A and B now yielded a t -statistic of 0.36 (n.s.), indicating (statistically) the absence of ability-related differences, on the average, between the two groups. That a nonsignificant difference in test performance was obtained when the SRE test data were corrected for chance success implies that the majority of examinees reached their difficulty ceiling within the time allowed, a first source of evidence in the present study that the time limit is adequate for deaf respondents.

Since the test score means of groups A and B were found not to differ significantly when formula scoring was applied to the data, groups A and B were combined. This combined data base was used to establish a reference distribution within which group C respondents' initial and revised measurements were compared for significant change. Table 3 summarizes the descriptive statistics resulting from this pooling and, additionally, includes an estimate of the reliability of the SRE in terms of internal consistency.

In order to examine the significance of changes in test scores when respondents in group C were allowed to finish the test and, so, to shed additional light on the adequacy of the SRE time limit, individuals' initial and revised measurements were compared in terms of the standard error of measurement (SEM) computed from the "reference" data (see Table 3). Specifically, to the extent that the observed changes in test performance are overshadowed by errors of measurement, it would be expected that group C respondents' "revised" test scores would lie within a confidence band extending ± 2 SEM about individuals' initial test scores. In a case by case analysis of subjects' test records, only five of the 36 persons included in group C (approximately 14%) deviated from this

Table 3
Summary Statistics from Pooled Data

	Raw Scores	Formula Scores
$\bar{x}$	37.65	31.34
sd	11.66	15.31
s _{$\bar{x}$}	.71	.93
r _{xx}	.93	.93
sem	2.99	3.91

expectation. Thus, all of the available statistical evidence from the present study indicates that, when formula scoring is applied to the test records of deaf students tested with the SRE, the recommended time limit poses no threat to measurement validity.

Abstract Reasoning Test (ABR)

Of the 308 persons tested with the ABR, 245 (79.5%) responded to all items (50 total) and completed the test within the time limit. For those individuals not completing the test within the time allowed ($n = 63$ or 20.5% of the study sample), the mean number of items for which responses were not provided was approximately six (range 1-26). However, this distribution (i.e., number of items not reached) was highly skewed, with more than half of those respondents unable to complete the test having three or fewer items remaining when the time limit had expired. Generally, then, the time factor was less an issue with the ABR than was the case with the SRE.

As with the analyses performed on the SRE data, the ABR study sample was partitioned into three groups of respondents. Group A consisted of the 245 examinees who had completed the test within the time allowed. Group B consisted of 52 individuals who had not completed the test, and were not given the opportunity to do so. As before, group C consisted of those individuals (11 total) who initially did not complete the test, but were allowed to finish.

Table 4 provides descriptive statistics summarizing the test performance of the three groups of examinees. As in the previous analysis of the SRE data, the test score means of groups B and C did not differ significantly ($t = 0.21$, n.s.), while the difference between the means for groups A and B reached statistical significance ($t = 2.95$, $p < .05$).

Table 5 summarizes the test performance of the three groups of examinees after the correction for chance formula was applied to all individuals' test data. Unlike the analysis of the SRE data, however, the difference between the means of groups A and B remained

Table 4
Test Performance of Three Groups of Examinees (ABR)

	Group A	Group B	Group C
n	245	52	11
$\bar{x}$	35.49	30.75	30.09
sd	9.06	10.78	8.53
$s_{\bar{x}}$	.58	1.50	2.70

statistically significant ($t = 2.37, p < .05$) suggesting, in this instance, the operation of ability related factors, over and above the propensity toward guessing, as a determinant in test completion. Thus, it is suggestive that respondents who completed the ABR found the test less difficult than did respondents who did not complete the test, in which case an extension of testing time for non-completers would seem to serve no useful purpose (i.e., their ceilings of difficulty already having been reached).

The finding of statistical difference above cautions against the pooling of data for groups A and B. Therefore, only group A data (complete test records) were utilized in the calculation of the SEM for the ABR. Based upon these 245 test records, the SEM was estimated at 3.24 (Hoyt $r_{xx} = .92$).

Using a confidence interval spanning ± 2 SEM about individuals' initial timed test scores, it would be predicted (under the null hypothesis) that only one of the 11 persons included in group C would attain a revised test score significantly higher than the initial estimate. In fact, this is precisely what was observed in case analysis of the group C data. Consequently, no evidence was found to indicate that the present timed condition for the ABR is biased against deaf respondents.

CONCLUSIONS

Spearman (1927) differentiated experimentally between two types of speed factors which enter into performance on mental tests administered under timed conditions. One of these factors involves the speed of mental operations and is reflected in the time required for an individual to recall relevant information or, otherwise, to manipulate elements of a problem in arriving at a solution. The other form of speed was described by Spearman as attitudinal in nature, involving a general preference for speed in performing tasks, and has sometimes been referred to as personal tempo (Jensen, 1980).

While a timed test would almost certainly seem to penalize persons with slower cognitive speed or personal tempo, no evidence was adduced in the present study to

indicate that deaf examinees' scores on the tests studied increased appreciably when the time limits were extended. This is in part due to the manner in which the tests were constructed (i.e., spiral omnibus), allowing the majority of examinees to reach their ceiling of difficulty within the time allowed. However, the results of the present study rest, in part, on the assumption that examinees' performance under the timed and untimed conditions is directly comparable.

In summary, data obtained in the present study failed to indicate that the time limits observed in the administration of the DAT subtests studied (i.e., Space Relations and Abstract Reasoning) threatens the validity of the measurements obtained for deaf respondents. However, this conclusion remains instrument-specific, insofar as the measures studied contain a low verbal component. Educational and psychological tests that pose greater demands on verbal language acquisition have been regarded as problematic owing to the language limitations imposed by deafness (Vernon, 1980) and, therefore, may require special procedural conditions for their administration. Consequently, generalization of the findings presented in this paper to other testing contexts remains uncertain.

REFERENCES

- Bennett, G. K., Seashore, H. G., & Wesman, A. G. Differential Aptitude Tests. New York: Psychological Corporation, 1966.
- Cronbach, L. J. Essentials of Psychological Testing (2nd ed.). New York: Harper, 1960.
- Jensen, A. R. Bias in Mental Testing. New York: The Free Press, 1980.
- Spearman, C. The Abilities of Man. New York: Macmillan, 1927.
- Stanley, J. C., & Hopkins, K. D. Educational and psychological measurement and evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1972.
- Vernon, M. Vocational assessment: Approaches and implications. Directions, 1980, 1 (4), 31-35.

Footnotes

¹For this and subsequent t-tests reported in this paper, with only one exception, the homogeneity of variance assumption was found to be tenable using the F test. The variance terms for groups A and B reported in Table 2 were found to differ significantly ($p < .05$); therefore, in this instance, the separate variance estimate was used rather than the pooled or average variance to compute a t value.

²While a variety of correction formulae exist, the expression most frequently used (and the one utilized in the present study) is as follows:

$$\text{Score} = R - \frac{W}{n-1}$$

where

R = number of correct responses,

W = number of incorrect responses,

n = number of response options per item.

Omitted items are not reflected in W.